TRANSMITTAL LETTER OT THE UNITED STATES DESIGNATED/ELECTED OFFICE (DO/EO/US) CONCERNING A FILING UNDER 35 U.S.C. 371		Attorney Docket No. **0512-1024** U.S. Application No. 10/088895
INTERNATIONAL APPLN. NO. **PCT/FR00/02640**	INTERNATIONAL FILING DATE **22 SEPTEMBER 2000 (22.09.00)**	PRIORITY DATE CLAIMED **24 SEPTEMBER 1999 (24.09.99)**

TITLE OF INVENTION: **METHOD OF THEMATICALLY CLASSIFYING DOCUMENTS, A THEMATIC CLASSIFICATIN MODULE, AND A SEARCH ENGINE INCORPORATING SUCH A MODULE**

APPLICANT(S) FOR DE/EO/US: **LAURENT BIETTRON, FRÉDERIC PALLU AND SYLVIE TRICOT**

Applicant herewith submits to the United States Designated Elected Office (DO/EO/US) the following items and other information:

1. ☒ This is a FIRST submission of items concerning a filing under 35 U.S.C. 371.
2. ☐ This is a SECOND or SUBSEQUENT submission of items concerning a filing under 35 U.S.C. 371.
3. ☒ This is an express request to begin national examination procedures (35 U.S.C. 371(f)). The submission must include items (5), (6), (9) and (21) indicated below.
4. ☒ The US has been elected by the expiration of 19 months from the priority date (Article 31).
5. ☒ A copy of the International Application as filed (35 U.S.C. 371 (c)(2))
   a. ☒ is attached hereto (required only if not communicated by the International Bureau)
   b. ☐ has been communicated by the International Bureau. See attached PCT/IB/308.
   c. ☐ is not required, as the application was filed in the United States Receiving Office (RO/US).
6. ☒ An English language translation of the International Application as filed (35 U.S.C. 371 (c)(2))
   a. ☒ is attached hereto.
   b. ☐ has been previously submitted under 35 U.S.C. 154(d)(4).
7. ☐ Amendments to the claims of the International Application under PCT Article 19 (35 U.S.C. 371 (c)(3))
   a. ☐ are attached hereto (required only if not communicated by the International Bureau).
   b. ☐ have been communicated by the International Bureau.
   c. ☐ have not been made, however, the time limit for making such amendments has NOT expired.
   d. ☐ have not been made and will not be made.
8. ☐ An English language translation of the amendments to the claims under PCT Article 19 (35 U.S.C. 371 (c)(3)).
9. ☐ An oath or declaration of the inventor(s) (35 U.S.C. 371(c)(4)).
10. ☐ An English language translation of the annexes of the International Preliminary Examination Report under PCT Article 36 (35 U.S.C. 371(c)(5)).

Items 11 to 20 below concern document(s) or information included:

11. ☒ Information Disclosure Statement (IDS) w/PTO-1449 - ☒ Copy of IDS citations
12. ☐ Assignment Papers (cover sheet & document(s))
13. ☒ A FIRST Preliminary Amendment.
14. ☐ A SECOND or SUBSEQUENT Preliminary Amendment.
15. ☐ A substitute specification.
16. ☐ A change of power of attorney and/or address letter.
17. ☐ A computer-readable form of the sequence listing in accordance with PCT Rule
18. ☐ A second copy of the published international application under 35 U.S.C. 154(d)(4).
19. ☐ A second copy of the English language translation of the international application (35 U.S.C. 154(d)(4)).
20. ☒ Other items or information: **INTERNATIONAL PRELIMINARY EXAMINATION REPORT (PCT/IPEA/409), INTERNATIONAL SEARCH REPORT (PCT/ISA/210), APPLICATION DATA SHEET, ABSTRACT**

U.S. APPLICATION NO.	INTERNATIONAL APPLN. NO.	ATTORNEY DOCKET NO.
10/088895	PCT/FR00/02640	0512-1024

				CALCULATIONS	PTO USE ONLY
21. ☒ The following fees are submitted: BASIC NATIONAL FEE (37 CFR 1.492 (a) (1)-(5): Neither international preliminary examination fee nor international search fee paid to USPTO and international Search Report not prepared by the EPO or JPO ....$1040.00 International preliminary examination fee not paid to USPTO but International Search Report prepared by the EPO or JPO ....$890.00 International preliminary examination fee not paid to USPTO but International search fee paid to USPTO ....$740.00 International preliminary examination fee paid to USPTO but all claims did not satisfy provision of PCT Article 33 (1)-(4) ....$710.00 International preliminary examination fee paid to USPTO and all claims satisfied provision of PCT Article 33 (1)-(4) $100.00					
ENTER APPROPRIATE BASIC FEE AMOUNT				$ 890.00	
Surcharge of $130.00 for furnishing the oath or declaration later than □ 20- ☒ 30 months from the earliest claimed priority date (37 CFR 1.492(e))				$ 130.00	
CLAIMS	NUMBER FILED	NUMBER EXTRA	RATE		
Total Claims	13 - 20 =	0	X $18.00	$	
Independent Claims	2 - 3 =	0	X $84.00	$	
MULTIPLE DEPEND CLAIM(S) (if applicable)			+ $280.00	$	
TOTAL OF ABOVE CALCULATION -				$ **1020.00**	
□ Applicant claims small entity status. See 37 CFR 1.27. The fees indicated above are reduced by ½. +				$	
SUBTOTAL =				$ 1020.00	
Processing fee of $130.00 for furnishing the English translation later than □ 20 □ 30 months from the earliest claimed priority date (37 CFR 1.492Z(f)).				$	
TOTAL NATIONAL FEE =				$ 1020.00	
Fee for recording the enclosed assigned (37 CFR 1.21(h)). The assignment must be accompanied by an appropriate cover sheet (37 CFR 3.28, 3.31) $40.00 per property +				$	
TOTAL FEES ENCLOSED -				$ 1020.00	
				Amount to be refunded:	$
				Charged:	$

☒ A Check in the amount of **$1,020.00** to cover all fees is attached.

□ The Commissioner is hereby authorized to charge indicated fees and credit any overpayments to Deposit account No. 25-0120 in the name of Young & Thompson, as described below. A duplicate copy of this sheet is enclosed.

☒ The Commissioner is hereby authorized in this, concurrent, and future replies, to charge payment or credit any overpayment to Deposit Account No. 25-0120 for any additional fee required under 37 C.F.R. §§ 1.16 or 1.17.

SEND ALL CORRESPONDENCE TO:
745 South 23rd Street
Arlington, VA 22202
Telephone (703) 521-2297
Y&T Customer No. 000466

00466
PATENT TRADEMARK OFFICE

Benoit Castel
SIGNATURE

Benoit Castel
NAME

35,041
REGISTRATION NO.

BC/bam
Date: **March 25, 2002**

PATENT
0512-1024

IN THE U.S. PATENT AND TRADEMARK OFFICE

In re application of: Laurent BIETTRON et al.

Appl. No.: **NEW** Group:

Filed: March 25, 2002 Examiner:

For: METHOD OF THEMATICALLY CLASSIFYING DOCUMENTS, A THEMATIC CLASSIFICATION MODULE, AND A SEARCH ENGINE INCORPORTING SUCH A MODULE

PRELIMINARY AMENDMENT

Assistant Commissioner for Patents
Washington, DC 20231

March 25, 2002

Sir:

The following preliminary amendments and remarks are respectfully submitted in connection with the above-identified application.

IN THE ABSTRACT OF THE DISCLOSURE:

Delete the Abstract as originally filed which appears on the cover page of the Published Application. Add new Abstract as enclosed herewith on a separate sheet.

IN THE SPECIFICATION:

Please replace the paragraph beginning on page 2, line 4, with the following rewritten paragraph:

--Manually classifying document pages involves high creation and updating costs while allowing only a limited number of pages to be indexed. Consequently, some requests do not obtain any response.--

Please add the following paragraph before the paragraph beginning on page 2, line 4:

--Another method described in document US-A-5 625 767 enables thermatic classification to be performed on the basis of a statistical analysis of the document. However, that method requires the documents to be manually classifed beforehand.--

<u>IN THE CLAIMS</u>:

Please cancel claims 1-13 without prejudice or disclaimer of the subject matter contained therein.

Please add the following claims:

--14/(New) A method of thematically classifying documents, in particular for making up or updating thematic databases for a search engine, the method comprising the following steps:

- manually and/or automatically selecting a sample of documents representative of each theme;

- automatically identifying within the selected documents elements that are characteristic of each theme;

- automatically allocating a coefficient to each identified element, which coefficient is representative of the relevance of said element relative to the corresponding theme;

- for each document to be classified, identifying said theme-characterizing elements that are contained in the document for each of the themes, and for each theme corresponding to the elements, using the coefficients allocated to said elements to calculate the value of a characteristic representative of the relevance of that theme for the document, in order to decide whether or not the document relates to the theme, said identification and calculation steps being performed automatically for each document downloaded from a computer network;

- automatically classifying the downloaded documents as a function of the themes with which they deal; and

- automatically storing the documents classified thematically in databases that can be interrogated on the basis of themes contained in a request;

and the step of allocating said coefficient to each identified element comprises the following steps for each theme:

- automatically calculating the frequency of the element in the selected documents relating to the theme;

- automatically calculating the frequency of the element in the selected documents that do not relate to the theme; and

- automatically calculating the ratio of the calculated frequencies.

15/(New) A method according to claim 14, further comprising the step of automatically sorting themes in a theme tree structure in decreasing order of coefficients.

16/(New) A method according to claim 14, wherein the step of automatically calculating the characteristic representative of the relevance of the theme of a document for classification comprises the following steps, for each theme:

- reading the value of the ratio of said frequencies for each theme-representing element extracted from the document;

- multiplying together the values as read; and

- allocating the result of this multiplication to the value of said characteristic.

17/(New) A method according to claim 14, wherein it is automatically decided that the document relates to a theme if the value of said characteristic representative of the relevance of the theme for said document is greater than a threshold value.

18/(New) A method according to claim 17, wherein the threshold value for each theme is automatically determined on the basis of said frequency ratios using the following relationship:

$$\text{score - threshold}_{\text{theme}} = (R_{\text{mean}})\text{theme\_n}$$

in which:

score - $\text{threshold}_{\text{theme}}$ designates the threshold value;

R_{mean} represents the mean value of the frequency ratios R of the elements of the theme; and

theme_n designates a predetermined number.

19/(New) A method according to claim 17, wherein the threshold value is adjusted manually.

20/(New) A method according to claim 14, wherein the steps of automatically identifying theme-characterizing elements contained in a document and for each theme are performed by means of a hashing table.

21/(New) A method according to claim 14, wherein for each vocabulary element of a request formulated by a user, coefficients are automatically calculated characteristic of the element relative to each known theme, and each element is associated with the corresponding themes and coefficients, so that said coefficients reach a minimum value.

22/(New) A module for thematically classifying documents, in particular for a search engine, the module comprising a central processor unit having means for comparing elements extracted from each document with elements characteristic of various themes, each element being allocated a coefficient representative of the relevance of said element for a corresponding theme, and means for calculating the value of at least one characteristic representative of the relevance of a theme for the document on the basis of the coefficients of said characteristic elements that the document contains, in order to decide whether or not the document relates to said theme, said central unit being connected to means for storing documents classified by theme that can be interrogated on the basis of themes contained in a request, and the module has means for calculating the frequency of the element in the selected documents relating to the theme, means for calculating the frequency of the element in the selected documents that do not relate to the theme, and means for calculating the ratio between the calculated frequencies.

23/(New) The use of a module for thematically classifying documents according to claim 22 to determine which themes are contained in a request formulated by a user.

24/(New) The use of a module for thematically classifying documents according to claim 22 for determining which themes are contained in pages downloaded from a computer network or in a request formulated by a user, and for filtering downloaded documents to ban consultation of pages relating to one or more predetermined themes.

25/(New) The use of a module for thematically classifying documents according to claim 22 to determine which themes are contained in a request formulated by a user and for generating user profiles on the basis of the themes to which the request relates.

26/(New) A search engine for documents on a computer network, the engine comprising an indexing module for creating and updating thematic databases on the basis of documents downloaded from the computer network, and a module for interrogating thematic databases adapted to supply the references of documents corresponding to a request that has been input thereto, the search engine further comprising a thematic classification module according to claim 22 associated with the indexing module.--

REMARKS

Claims 14-26 are pending in the present application. Claims 1-13 have been cancelled and claims 14-26 have been added.

Entry of the above amendments is earnestly solicited. An early and favorable first action on the merits is earnestly requested.

Should there be any matters that need to be resolved in the present application, the Examiner is respectfully requested to contact the undersigned at the telephone number listed below.

Attached hereto is a marked-up version of the changes made to the specification by the current amendment. The attached page is captioned "VERSION WITH MARKINGS TO SHOW CHANGES MADE."

The Commissioner is hereby authorized in this, concurrent, and future replies, to charge payment or credit any overpayment to Deposit Account No. 25-0120 for any additional fees required under 37 C.F.R. § 1.16 or under 37 C.F.R. § 1.17.

Respectfully submitted,

YOUNG & THOMPSON

Benoit Castel

Benoit Castel, Reg. No. 35,041

745 South 23rd Street
Arlington, VA 22202
Telephone (703) 521-2297

BC/bam
Attachments

VERSION WITH MARKINGS TO SHOW CHANGES MADE

IN THE SPECIFICATION:

The paragraph beginning on page 2, line 4, has been amended as follows:

~~However manually~~ Manually classifying document pages involves high creation and updating costs while allowing only a limited number of pages to be indexed. Consequently, some requests do not obtain any response.

ABSTRACT

A METHOD OF THEMATICALLY CLASSIFYING DOCUMENTS, A THEMATIC CLASSIFICATION MODULE, AND A SEARCH ENGINE INCORPORATING SUCH A MODULE

This method of thematically classifying documents, in particular for making up or updating thematic databases for a search engine, comprises the steps of selecting documents representative of each theme, identifying within the selected documents, elements that are characteristic of each theme, allocating a coefficient to each identified element, the coefficient being representative of the relevance of the element relative to the corresponding theme, and for each document for classification, identifying the elements characteristic of each theme contained in the document and, for each theme corresponding thereto, using the coefficients allocated to the elements to calculate the value of a characteristic representative of the relevance of the theme for the document, in order to decide whether or not the document relates to the theme.

A METHOD OF THEMATICALLY CLASSIFYING DOCUMENTS, A THEMATIC CLASSIFICATION MODULE, AND A SEARCH ENGINE INCORPORATING SUCH A MODULE

The present invention relates to a method of thematically classifying documents and intended in particular for setting up or updating thematic databases, in particular for a search engine.

The invention also relates to a module for thematically classifying documents, and to a search engine fitted with such a thematic classification module.

At present, two main computer tools are known for searching documents on a computer network such as the Internet, for example.

These tools are search engines and guides.

A search engine is a tool that serves to extract the words or terms that are most representative of information, mainly in the form of text, and to store them in a database, also known as an "index" base.

Such index bases are generally updated relatively frequently.

In response to a request made by a user, the same tool scans through the index bases in order to identify the terms which are most relevant relative to those of the request, and then to sort the information obtained in return.

The other technique for searching for documents on a computer network consists in using a guide. That tool proposes searches by category, with document pages being classified manually by researchers.

Those types of tool present various drawbacks.

Firstly, search engines do not propose classifying document pages by category. The pages provided in response to a request are not typified. Thus, ambiguous requests can give rise to a very wide variety of responses that are perceived by the user as noise.

In contrast, guides provide a user with responses that are typified, i.e. that relate to the same theme(s) as the request.

However manually classifying document pages involves high creation and updating costs while allowing only a limited number of pages to be indexed. Consequently, some requests do not obtain any response.

The object of the invention is to mitigate the drawbacks of search engines and of guides.

The invention thus provides a method of thematically classifying documents, in particular for making up or updating thematic databases for a search engine, the method being characterized in that it comprises the following steps:

- selecting a sample of documents representative of each theme;
- identifying within the selected documents elements that are characteristic of each theme;
- allocating a coefficient to each identified element, which coefficient is representative of the relevance of said element relative to the corresponding theme;
- for each document to be classified, identifying said theme-characterizing elements that are contained in the document for each of the themes, and for each theme corresponding to the documents, using the coefficients allocated to said elements to calculate the value of a characteristic representative of the relevance of that theme for the document, in order to decide whether or not the document relates to the theme, said identification and calculation steps being performed automatically for each document downloaded from a computer network;
- classifying the downloaded documents as a function of the themes with which they deal; and
- storing the documents classified thematically in databases that can be interrogated on the basis of themes contained in a request;

and in that the step of allocating said coefficient to each identified element comprises the following steps for each theme:

- calculating the frequency of the element in the selected documents relating to the theme;
- calculating the frequency of the element in the selected documents that do not relate to the theme; and
- calculating the ratio of the calculated frequencies.

The documents downloaded from a computer network are thus classified as a function of the themes dealt with therein, and this is done automatically.

The classification method of the invention can also include one or more of the following characteristics, taken singly or in any technically feasible combination:

- it further comprises the step of sorting themes in a theme tree structure in decreasing order of coefficients;
- the step of calculating the characteristic representative of the relevance of the theme of a document for classification comprises the following steps, for each theme:
    - reading the value of the ratio of said frequencies for each theme-representing element extracted from the document;
    - multiplying together the values as read; and
    - allocating the result of this multiplication to the value of said characteristic;
- deciding that the document relates to a theme if the value of said characteristic representative of the relevance of the theme for said document is greater than a threshold value;
- determining the threshold value for each theme on the basis of said frequency ratios using the following relationship:

$$\text{score} - \text{threshold}_{\text{theme}} = (R_{\text{mean}})\text{theme\_n}$$

in which:

score - $\text{threshold}_{\text{theme}}$ designates the threshold value;

R_{mean} represents the mean value of the frequency ratios R of the elements of the theme; and

theme_n designates a predetermined number;

- in a variant, adjusting the threshold value manually;

- performing the steps of identifying theme-characterizing elements contained in a document for each theme by means of a hashing table; and

- for each vocabulary element of a request formulated by a user, calculating coefficients characteristic of the element relative to each known theme, and associating each element with the corresponding themes and coefficients, so that said coefficients reach a minimum value.

When searching index entries, i.e. while searching for documents that correspond to the request, it is also possible directly to access the themes associated with each element and the corresponding coefficients which are-combined by multiplication in order to determine a classification of themes associated with the entire request.

The invention also provides a module for thematically classifying documents, in particular for a search engine, the module being characterized in that it comprises a central processor unit having means for comparing elements extracted from each document with elements characteristic of various themes, each element being allocated a coefficient representative of the relevance of said element for a corresponding theme, and means for calculating the value of at least one characteristic representative of the relevance of a theme for the document on the basis of the coefficients of said characteristic elements that the document contains, in order to decide whether or not the document relates to said theme, said central unit being connected to means for storing documents classified by theme that can be interrogated on the basis of themes contained in a request, and in that the module has means for calculating the frequency of the element in the selected documents relating to the theme, means for calculating the frequency of the element in the selected documents that do not relate to the theme, and means for calculating the ratio between the calculated frequencies.

The invention also provides a search engine for documents on a computer network, the engine comprising an indexing module for creating and updating thematic databases on the basis of documents downloaded from the computer network, and a module for interrogating thematic databases adapted to supply the references of documents corresponding to a request that has been input thereto, the search engine being characterized in that it further comprises a thematic classification module as defined above associated with the indexing module.

Other characteristics and advantages appear from the following description given purely by way of example and made with reference to the accompanying drawings, in which:

- Figure 1 is a flow chart showing the main operating stages of a module of the invention for thematically classifying documents for a search engine;
- Figure 2 is a flow chart showing the method of calculating the elements characteristic of themes; and
- Figure 3 is a flow chart showing the method of calculating the themes of a document.

Figure 1 shows the main stages of the method of the invention for thematically classifying documents.

It is intended to enable documents downloaded from a computer network to be classified as a function of the themes they deal with. For example, it can be implemented within a search engine.

Under such circumstances, it is involved in the indexing process, and also during processing of a request formulated by a user so as to determine all of the themes dealt with in the request.

Nevertheless, it will be understood that other applications can be envisaged. For example, the method can be implemented at a network access point for stations using an Internet network in order to determine the nature of the web pages downloaded by the users and to filter requests in order to authorize or ban certain themes, for example themes contrary to *ordre public* or morals, or indeed to calculate statistics concerning uses' centers of interest.

To proceed with this classification, the method comprises two distinct stages, namely: a prior first stage of acquiring the thematic vocabulary of the corpus of documents and of giving each word of the vocabulary a threshold value above which it is decided that a document containing this word relates to the corresponding theme; and also a second stage of classification proper, during which a document downloaded from the network is automatically classified as a function of the characteristic elements it contains.

By way of example, this second stage takes place periodically, and only documents that have been newly created or modified are classified.

The first stage of thematic vocabulary acquisition is described below with reference to Figures 1 to 3.

As can be seen in Figure 1, this stage starts with a manual selection step 10 from a set 12 of samples (or "corpus") of documents that are representative of each of themes A to Z used for classifying documents during the second stage.

Thus, at the end of this manual selection step 10, a set of document corpuses such as 14 is available with each corpus relating to a particular theme (theme A, ..., theme Z). Naturally, the selection step can equally well be performed by any means other than manual.

During this selection step 10, a corpus 16 is also created of documents that do not relate to any of the themes A to Z, and a nomenclature 18 for the themes A to Z is defined, i.e. a list of said themes associated with subthemes relating thereto.

During the following step 20, these elements are input to a thematic classification module in order to extract from each document elements that are characteristic of each theme and to give each of them a coefficient representative of its relevance relative to a corresponding theme.

By way of example, this thematic classification module is in the form of a specific module of a search engine associated with an indexing module that creates or updates thematic databases.

It can also be implemented in the form of a specific module provided at an access point to a computer network, in particular an Internet network.

The module has software means suitable for extracting elements that are characteristic of each theme and for allocating respective coefficients representative of their relevance relative to the various themes, as described in detail below.

During this step 20, the classification module extracts the elements characteristic of each theme from each of the selected documents.

This extraction is performed using a computer tool of conventional type. It is therefore not described below.

At the end of this step 20, lists are available of elements that are characteristic of the themes A to Z, such as the lists 22.

With reference to Figure 2, this procedure of identifying the vocabulary that is characteristic of each theme is performed successively for each element extracted from the documents in each of the corpuses 14 and 16.

During a first step 24, a table of all candidate themes is cleared, i.e. a table of all themes that might correspond to an extracted element.

During the following step 26, a coefficient R is calculated for each theme, where the coefficient R is representative of the relevance of the element relative to the theme.

To proceed with this calculation, the frequency p of the element in the documents relating to the theme is initially calculated, and so is the frequency q of the same element in the documents that do not relate to the theme.

Thereafter the coefficient R is calculated which is constituted by the ratio of the frequencies p and q.

During the following step 28, a check is made to verify that the characteristics p, q, and R lie within predetermined limits.

If this is not the case, then the following element is processed.

If this is the case, then the theme is added to the table of candidate themes with a score equal to the coefficient R (step 30).

If any elements remain to be processed (step 32) then the procedure returns to preceding step 24.

Otherwise the procedure ends.

It will be observed that after the table of candidate themes has been filled it is preferably sorted by decreasing order of the scores R. It should also be observed that for each candidate theme, and up to some desired maximum number, a new element taken from the list of elements characteristic of said theme is added while remaining within the limit of a desired maximum number of the n best elements per theme selected as a function of their respective scores R.

With reference once more to Figure 1, during the following step 34, the thematic classification module proceeds by means of an appropriate algorithm automatically to calculate a threshold value corresponding to a minimum threshold to be reached in order to decide that a document containing an element characteristic of a theme does or does not relate to the theme.

To perform this calculation, the classification module begins by calculating the mean value R_{mean} of the ratios R of the characteristic elements of each theme (step 36).

Thereafter, it calculates the threshold value score - threshold_{theme} using the following relationship:

$$\text{score} - \text{threshold}_{theme} = (R_{mean})\text{theme\_n}$$

where theme_n designates a predetermined number which is selected to be equal to 5, for example, for most themes.

It can thus be seen in Figure 1 that after automatically calculating the scores to be reached, lists of elements that are characteristic of each of the themes A to Z are made available, such as the list 40, with each element being associated with a score to be reached, i.e. a threshold value beyond which it is considered that a document relates to the theme.

After this stage of acquiring thematic vocabulary, implemented using a corpus of documents representative of various themes, the second stage of thematic classification proper can be performed in order to make up thematic databases given overall numerical reference 42 from documents collected automatically from the computer network by robots such as 44.

These documents are input to the thematic classification module which also receives an indication of the theme nomenclature 18 and the elements available from the outcome of above-mentioned step 34. This module proceeds automatically to calculate the themes on which a document relates (step 46).

To do this, it has all of the software means required for implementing the above-mentioned operations.

With reference to Figure 3, at the end of a first step 48 of this procedure, the indexing module extracts from each document 50 downloaded by the robots 44 those elements that are characteristic of the themes it contains.

By way of example, this step is performed by using a hashing table to search quickly through the lists of characteristic elements for the elements contained in each document.

After these elements have been extracted, the elements characteristic of the themes contained in the list 40 are identified from amongst them.

For each identified element, the classification module then calculates a characteristic value representative of the relevance of each theme for the document, on the basis of the coefficients given to the element.

To do this, during the following step 52, a variable "theme_score", representative of the score of the document in a given theme is set at 1, and this is done for all of the themes.

Thereafter, for each element of the document, and for each theme in the tree structure of themes, if the element lies within the list of elements characteristic of the theme, then the score R is read, i.e. the value of the frequency ratio for each element, and the values read for the score R for each of the elements are multiplied together.

The result of this multiplication is then used as the value for the theme_score characteristic (step 54).

It is then decided that the themes recognized in document 50 are those for which the theme_score characteristic reaches or exceeds the score that is to be reached for these themes (step 56).

Thus, at the end of this procedure, a set 57 of themes is available to which the downloaded document 50 relates.

It will be understood that this procedure for automatically calculating the themes of documents downloaded by the robots 44 enables the indexing module of a search engine to classify these documents as a function of the themes dealt with and to build up the thematic databases 42.

Such a procedure for automatically calculating document themes can also be used for determining which themes are dealt with in requests made by users.

To do this, starting from a request, for each of the elements of the interrogation vocabulary used in the request, the coefficients characteristic of said element relative to each of the known themes are calculated and each of these elements is associated with the coefficients and themes in such a manner that the coefficients reach a minimum value.

When searching for index entries corresponding to the elements of a request, i.e. in order to calculate the results, it is thus possible to access directly the theme which is associated with the elements and also their coefficients, and these are combined by multiplication using the same procedure as that described above in order to classify the themes associated with the request as a whole.

It can thus be understood that this procedure makes it possible to ask a user to refine a request, for example when the request is formulated in vague manner.

It will also be understood that this procedure which enables the themes contained in a request to be identified makes it possible to monitor user requests in order to establish statistics for defining user profiles as a function of requests.

It will thus be understood that the invention as described above can be used for searching for themes contained in pages downloaded from a computer network, for determining the themes contained in a request formulated by a user, and on the basis of such determination, for filtering requests and also downloaded pages in order to ban the formulation of requests or the downloading of pages relating to predetermined banned themes, and also to generate user profiles.

Nevertheless, it should be observed that in the context of determining themes contained in a request, the request is considered as constituting a document input to the thematic classification module of the invention.

The invention is not limited to the implementation described.

In a variant, it is also possible manually to adjust the value of the threshold from which it is decided that a document does or does not bear on a given theme.

CLAIMS

1/ A method of thematically classifying documents, in particular for making up or updating thematic databases for a search engine, the method being characterized in that it comprises the following steps:

- selecting a sample of documents representative of each theme;

- identifying within the selected documents elements that are characteristic of each theme;

- allocating a coefficient (R) to each identified element, which coefficient is representative of the relevance of said element relative to the corresponding theme;

- for each document (50) to be classified, identifying said theme-characterizing elements that are contained in the document for each of the themes, and for each theme corresponding to the elements, using the coefficients allocated to said elements to calculate the value of a characteristic representative of the relevance of that theme for the document (50), in order to decide whether or not the document relates to the theme, said identification and calculation steps being performed automatically for each document downloaded from a computer network;

- classifying the downloaded documents as a function of the themes with which they deal; and

- storing the documents classified thematically in databases that can be interrogated on the basis of themes contained in a request;

and in that the step of allocating said coefficient to each identified element comprises the following steps for each theme:

- calculating the frequency of the element in the selected documents relating to the theme;

- calculating the frequency of the element in the selected documents that do not relate to the theme; and

- calculating the ratio of the calculated frequencies.

2/ A method according to claim 1, characterized in that it further comprises the step of sorting themes in a theme tree structure in decreasing order of coefficients.

3/ A method according to claim 1 or claim 2, characterized in that the step of calculating the characteristic representative of the relevance of the theme of a document for classification comprises the following steps, for each theme:

- reading the value of the ratio (R) of said frequencies for each theme-representing element extracted from the document;
- multiplying together the values as read; and
- allocating the result of this multiplication to the value of said characteristic.

4/ A method according to any one of claims 1 to 3, characterized in that it is decided that the document relates to a theme if the value of said characteristic representative of the relevance of the theme for said document is greater than a threshold value.

5/ A method according to claim 4, characterized in that the threshold value for each theme is determined on the basis of said frequency ratios using the following relationship:

$$\text{score - threshold}_{\text{theme}} = (R_{\text{mean}})\text{theme\_n}$$

in which:

score - $\text{threshold}_{\text{theme}}$ designates the threshold value;

R_{mean} represents the mean value of the frequency ratios R of the elements of the theme; and

theme_n designates a predetermined number.

6/ A method according to claim 4, characterized in that the threshold value is adjusted manually.

7/ A method according to any one of claims 1 to 6, characterized in that the steps of identifying theme-characterizing elements contained in a document (50) and for each theme are performed by means of a hashing table.

8/ A method according to any one of claims 1 to 7, characterized in that for each vocabulary element of a request formulated by a user, coefficients are calculated characteristic of the element relative to each known theme, and each element is associated with the corresponding themes and coefficients, so that said coefficients reach a minimum value.

9/ A module for thematically classifying documents (50), in particular for a search engine, the module being characterized in that it comprises a central processor unit having means for comparing elements extracted from each document with elements characteristic of various themes, each element being allocated a coefficient (R) representative of the relevance of said element for a corresponding theme, and means for calculating the value of at least one characteristic representative of the relevance of a theme for the document on the basis of the coefficients of said characteristic elements that the document contains, in order to decide whether or not the document (50) relates to said theme, said central unit being connected to means for storing documents classified by theme that can be interrogated on the basis of themes contained in a request, and in that the module has means for calculating the frequency of the element in the selected documents relating to the theme, means for calculating the frequency of the element in the selected documents that do not relate to the theme, and means for calculating the ratio between the calculated frequencies.

10/ The use of a module for thematically classifying documents according to claim 9 to determine which themes are contained in a request formulated by a user.

11/ The use of a module for thematically classifying documents according to claim 9 for determining which themes are contained in pages downloaded from a computer network or in a request formulated by a user, and for filtering downloaded documents to ban consultation of pages relating to one or more predetermined themes.

12/ The use of a module for thematically classifying documents according to claim 9 to determine which themes are contained in a request formulated by a user and for generating user profiles on the basis of the themes to which the request relates.

13/ A search engine for documents on a computer network, the engine comprising an indexing module for creating and updating thematic databases on the basis of documents downloaded from the computer network, and a module for interrogating thematic databases adapted to supply the references of documents corresponding to a request that has been input thereto, the search engine being characterized in that it further comprises a thematic classification module according to claim 9 associated with the indexing module.

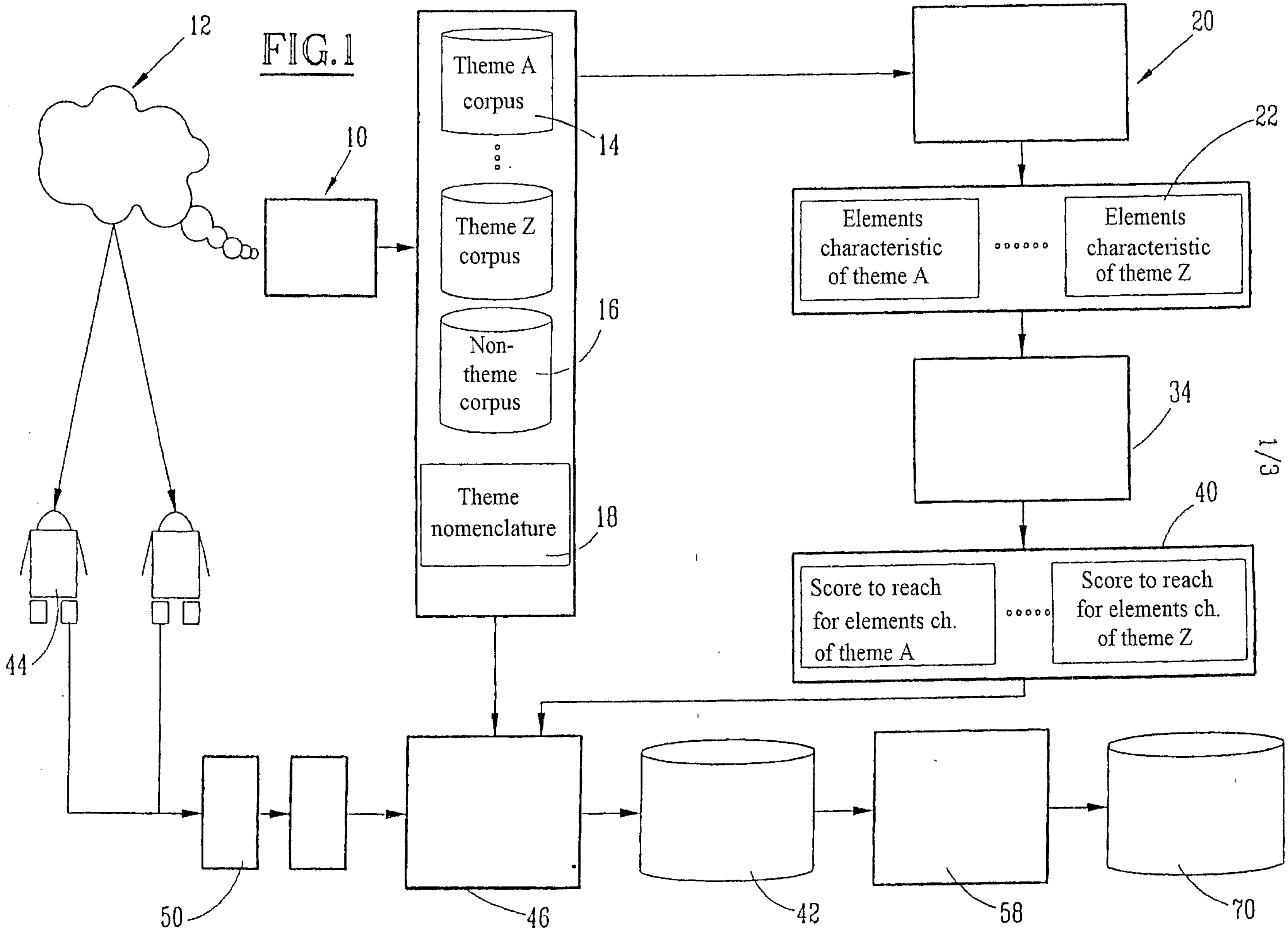
ABSTRACT

A METHOD OF THEMATICALLY CLASSIFYING DOCUMENTS, A THEMATIC CLASSIFICATION MODULE, AND A SEARCH ENGINE INCORPORATING SUCH A MODULE

This method of thematically classifying documents, in particular for making up or updating thematic databases (42) for a search engine, comprises the steps of selecting documents representative of each theme, identifying within the selected documents, elements that are characteristic of each theme, allocating a coefficient (R) to each identified element, said coefficient being representative of the relevance of said element relative to the corresponding theme, and for each document (50) for classification, identifying said elements characteristic of each theme contained in the document and, for each theme corresponding thereto, using the coefficients allocated to said elements to calculate the value of a characteristic representative of the relevance of the theme for the document (50), in order to decide whether or not the document relates to the theme.

Translation of the title and the abstract as they were when originally filed by the Applicant. No account has been taken of any changes that may have been made subsequently by the PCT Authorities acting ex officio, e.g. under PCT Rules 37.2, 38.2, and/or 48.3.

FIG.1


12
10
Theme A corpus
14
Theme Z corpus
Non-theme corpus
16
Theme nomenclature
18
20
22
Elements characteristic of theme A
Elements characteristic of theme Z
34
40
Score to reach for elements ch. of theme A
Score to reach for elements ch. of theme Z
44
50
46
42
58
70

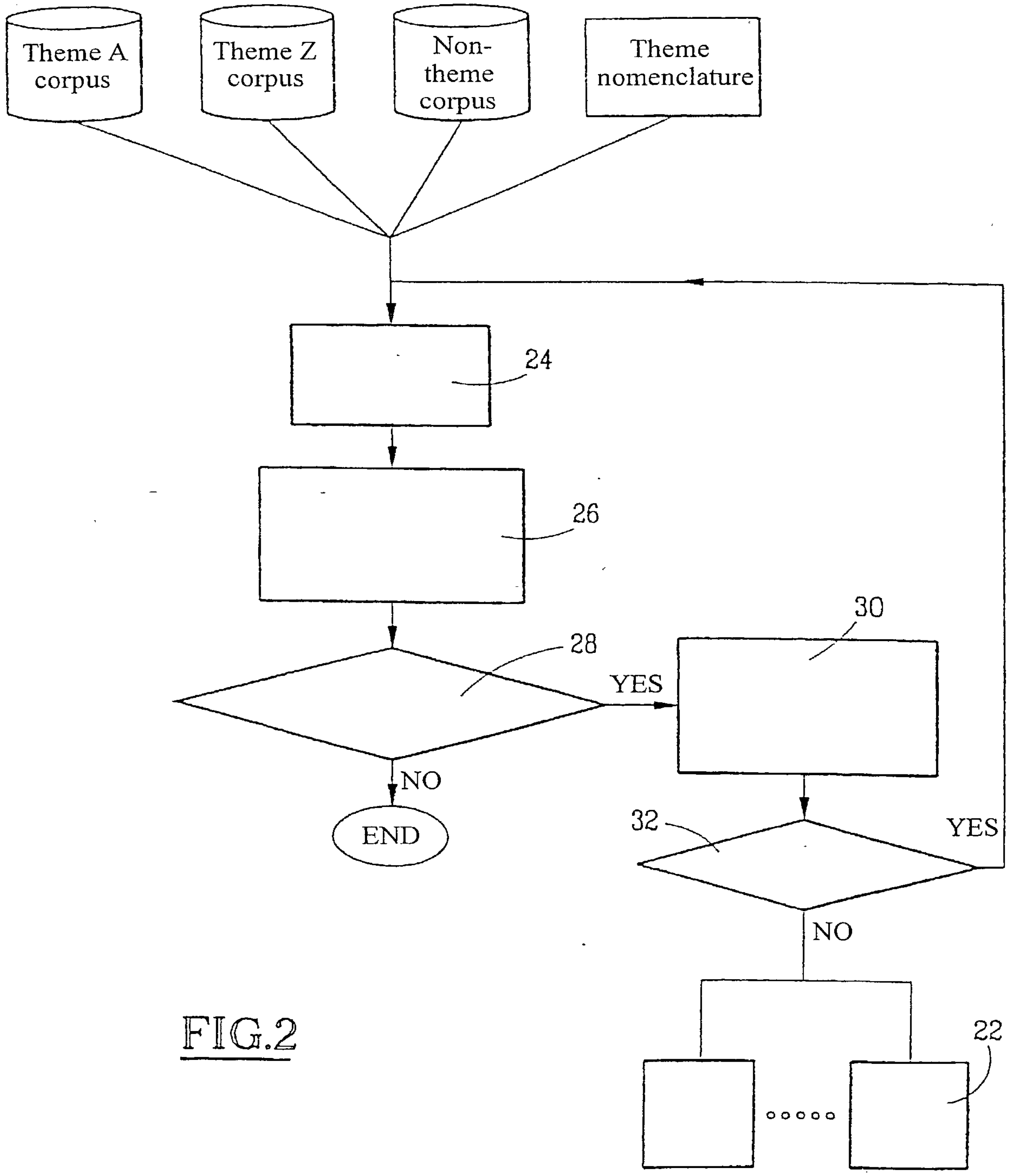

Theme A corpus
Theme Z corpus
Non-theme corpus
Theme nomenclature
24
26
28
YES
30
NO
END
32
YES
NO
22
ooooo

FIG.2

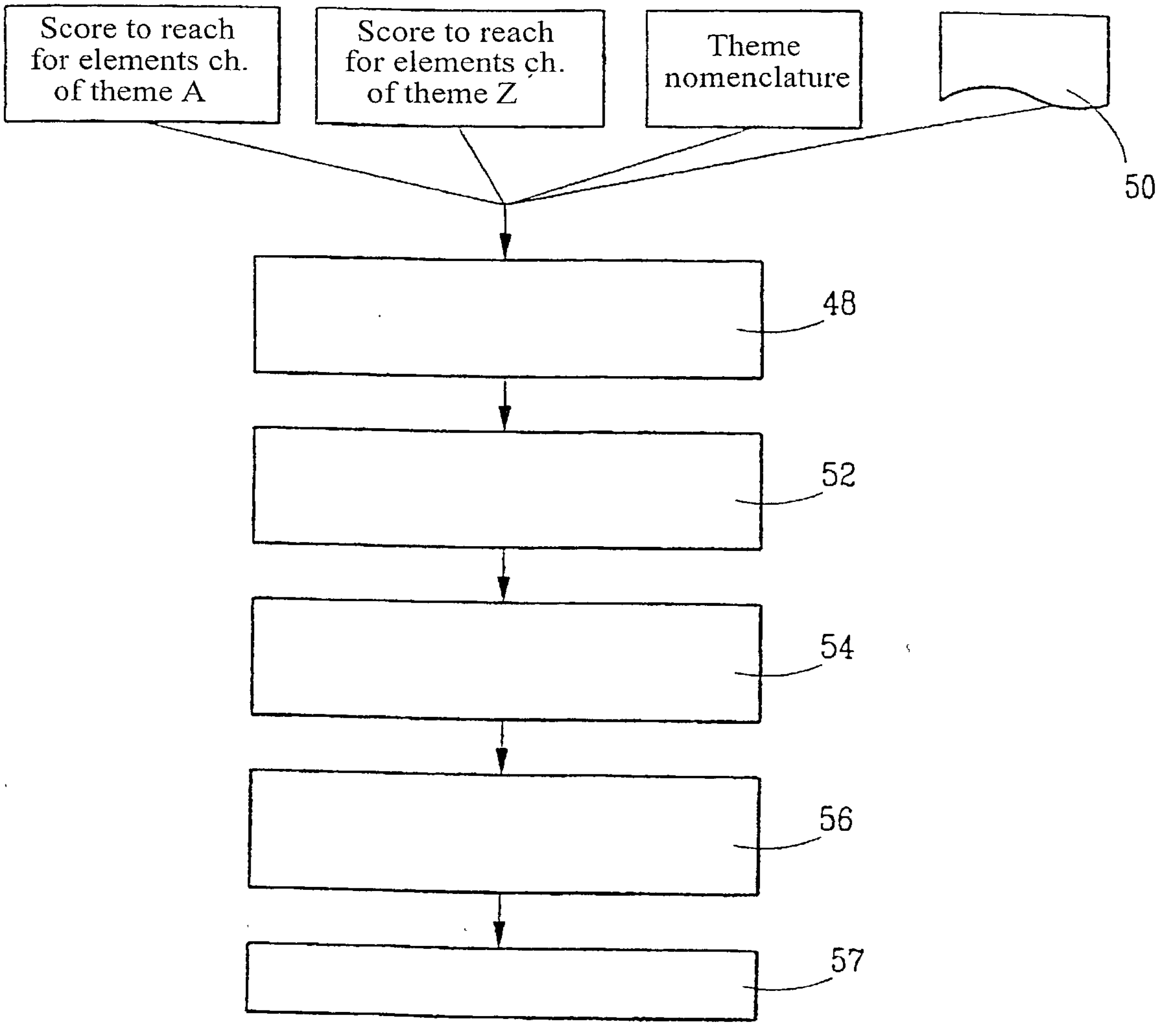

Score to reach for elements ch. of theme A
Score to reach for elements ch. of theme Z
Theme nomenclature
50
48
52
54
56
57

FIG.3

Ref. ______________

COMBINED DECLARATION AND POWER OF ATTORNEY

As a below named inventor, I hereby declare that

My residence, post office address and citizenship are as stated below next to my name.

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

A METHOD OF THEMATICALLY CLASSIFYING DOCUMENTS, A THEMATIC CLASSIFICATION MODULE, AND A SEARCH ENGINE INCORPORATING SUCH A MODULE

the specification of which: *(check one)*

REGULAR OR DESIGN APPLICATION

[ ] is attached hereto.

[ ] was filed on __________________ as application Serial No. ______________ and was amended on ________________ (if applicable).

PCT FILED APPLICATION ENTERING NATIONAL STAGE

[X] was described and claimed in International application No. PCT/FR0002640 filed on SEPTEMBER 22, 2000 and as amended on ____________________ (if any).

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to patentability as defined in Title 37, Code of Federal Regulations, §1.56.

PRIORITY CLAIM

I hereby claim foreign priority benefits under 35 USC 119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed.

PRIOR FOREIGN APPLICATION(S)

Country	Application Number	Date of Filing (day, month, year)	Priority Claimed
FRANCE	9911973	24.09.99	YES

(Complete this part only if this is a continuing application.)

I hereby claim the benefit under 35 USC 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of 35 USC 112, I acknowledge the duty to disclose information which is material to patentability as defined in Title 37 Code of Federal Regulations §1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application:

(Application Serial No.) (Filing Date) (Status--patented, pending, abandoned)

POWER OF ATTORNEY

The undersigned hereby authorizes the U.S. attorney or agent named herein to accept and follow instructions from ________________ as to any action to be taken in the Patent and Trademark Office regarding this application without direct communication between the U.S. attorney or agent and the undersigned. In the event of a change in the persons from whom instructions may be taken, the U.S. attorney or agent named herein will be so notified by the undersigned.

As a named inventor, I hereby appoint the registered patent attorneys represented by Customer No. **000466** to prosecute this application and transact all business in the Patent and Trademark Office connected therewith, including: **Robert J. PATCH, Reg. No. 17,355, Andrew J. PATCH, Reg. No. 32,925, Robert F. HARGEST, Reg. No. 25,590, Benoît CASTEL, Reg. No. 35,041, Eric JENSEN, Reg. No. 37,855, Thomas W. PERKINS, Reg. No. 33,027, and Roland E. LONG, Jr., Reg. No. 41,949,**

c/o YOUNG & THOMPSON,
Second Floor,
745 South 23rd Street,
Arlington, Virginia 22202.

00466
PATENT TRADEMARK OFFICE

Address all telephone calls to Young & Thompson at 703/521-2297. Telefax: 703/685-0573.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

1-ω Full name of sole or first inventor: Laurent BIETTRON
(given name, family name)

Inventor's signature ______________ Date 26/03/2002

Residence: 67 route de la Côte - 22300 LANNION - FRANCE FRV Citizenship: FRENCH

Post Office Address: The same as above

2-ω Full name of second joint inventor, if any: Frédéric PALLU
(given name, family name)

Inventor's signature F. PALLU Date 27/3/2002

Residence: Keravel - 22560 TREBEURDEN - FRANCE FRV Citizenship: FRENCH

Post Office Address: The same as above

3-ω Full name of third joint inventor, if any: Sylvie TRICOT
(given name, family name)

Inventor's signature S. Tricot Date 2/4/2002

Residence: 14, Hent Lann - 22300 TREDREZ -FRANCE FRV Citizenship: FRENCH

Post Office Address: The same as above

Full name of fourth joint inventor:
(given name, family name)

Inventor's signature ______________ Date ______________

Residence: Citizenship: